DOCUMENT RESUME

ED 159 895 FL 009 711

AUTHOR Tommola, Jorma
TITLE Testing Listening Comprehension Through Redundancy Reduction. Language Centre News, No. 1.
INSTITUTION Jyvaskyla Univ. (Finland). Language Center.
PUB DATE 78
NOTE 22p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Cloze Procedure; Cognitive Processes; Communication (Thought Transfer); Language Ability; *Language Instruction; *Language Tests; Linguistic Competence; Linguistic Theory; *Listening Comprehension; *Listening Tests; Memory; Perception Tests; Psycholinguistics; Recall (Psychological); *Redundancy; *Second Language Learning; Test Construction; Testing; Testing Problems; Test Interpretation

ABSTRACT

The idea of reducing the redundancy of a verbal message in a statistical way is presented as a practiced technique of language testing. Considering the temporality of speech comprehension, and the necessarily sequential intake of information, these cues may include the serial order of elements and transitional probability. To give the background of reduced redundancy tests, the contructivist view of listening comprehension is outlined as a creative, active cognitive operation with several implications: (1) it means that processing is facilitated by the linguistic and pragmatic organization of the message, together with its presentation in context; (2) it implies that memory, especially short-term storage, is an essential part of comprehension; and (3) it implies that the native listener is only partly bound by the properties of the signals he receives. Two reduced redundancy techniques of listening comprehension testing are reviewed that present the learner with messages that do not contain all the information they normally carry. In the noise test and the aural cloze test, the learner needs to mobilize his total awareness of the linguistic and pragmatic structure. Preliminary observations on an experimental test of aural cloze with Finnish learners imply that the tests have instructional value and may have a stronger theoretical basis than the completion tasks traditionally used to measure listening for details. (MCR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159895

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Jorma Tommola
Department of English
University of Turku

TESTING LISTENING COMPREHENSION THROUGH REDUNDANCY REDUCTION

1. Introduction

The idea of reducing the redundancy of a verbal message in a statistical way is by no means a novelty in language testing. It has been used at least since information theory began to have an effect on readability measurement, and, in consequence, on the measurement of individual differences in native-language reading comprehension skill (Taylor 1957; Bormuth 1962). In language description, information theory was reflected in the efforts to characterize language in terms of transitional probability - i.e. the likelihood that a given language item will be followed by certain others. Such models of language structure can now be shown to be inadequate in comparison with the hierarchical elegance of generative descriptions. Yet it seems that some of the testing techniques in current use are based, in part at least, on just these probabilistic views (witness the widespread interest in gap-filling).

One possible reason for this is the recent trend in the study of language performance, which emphasises the dimension of time and the dynamic use of knowledge in perception and production (Marslen-Wilson 1976; Wold 1976). Even though the language user's competence is probably best described as an atemporal hierarchical system, his performance always involves a serial event spread over time. In perception, this temporal sequence, the succession of speech signals, is translated into an atemporal plan or perceptual schema. In production, the direction is the opposite. Obviously, atemporal competence is the basis of performance. The listener is not, however, "a transformational linguist trying to map a sentence onto a grammatical theory" (Marslen-Wilson 1976:227). Therefore his performance may also be based on a heuristic utilisation of whatever cues he may find in the signal. Considering the temporality of speech comprehension, and the necessarily sequential

FL009911

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Language Centre for Finnish Universities

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

intake of information, these cues may well include the serial order of elements and transitional probability. Perhaps it is this fact that makes for the underlying connection between some present-day language tests and such concepts as information, redundancy, probability, and predictability.

In this paper, an attempt is first made to outline comprehension as a creative, constructive activity, and to discuss some of the implications of this view, especially as regards redundancy utilisation. Next, two reduced redundancy techniques of listening comprehension testing are reviewed. Finally, a few preliminary observations on an experimental reduced redundancy listening comprehension test are offered.

2. The background to reduced redundancy tests: the constructivist view of listening comprehension and some of its implications

According to the "constructivist" view, speech comprehension is an active cognitive operation. The sensory data, received during a fast and holistic stage variously referred to as "feature detection" or "sensation" (Massaro 1975), "sensing" (Rivers 1975), or the "preattentive stage" (Neisser 1966), are stored for a brief period in an echoic memory structure. The constructive process begins immediately with the synthesising of perceptual units (e.g. syllables) from the sensory data. Those portions of the speech wave that fail to fit into the auditory synthesis are rejected as irrelevant noise, and do not come to the listener's attention. The process continues with an active transformation and reduction of the information into a verbal, meaningful form. This transformation is made possible by the store of linguistic information in the listener's long-term memory. The resulting string of recognised words is stored in the short-term "working" memory. Rehearsal, which has been described by Rivers (1971:128) as "the recirculating of material through our cognitive system", is necessary before the meaning of the message can be derived: it keeps the perceived string of words 'alive' in order that they may be related to other parts of the utterance through a further process of recoding. In this further abstraction of the information, words are combined into phrases, phrases into clauses, and so on. Recoding - "chunking", as Miller (1956) calls it - is thus

assumed to be the process through which the meanings of larger units become available to the listener. It can be thought of as a repeated process that reduces, say, a string of words into successively more abstract, meaningful items ("chunks") in the working memory. The derivation of the meaning of longer stretches of speech would not be possible without this restructuring, since the capacity of the working memory is limited. As a result of the rehearsal/recoding - working memory loop, information is finally placed in long-term storage in forms radically different from those in which it was originally perceived. Sentences, for example, lose their surface characteristics, and their meaning is stored in some kind of deep-structure form (Neisser 1966; Rivers 1975; Wanner 1974; Massaro 1975).

Speech comprehension is a communicative act, and communication takes place in real time. Because of this temporality, it seems that models attempting to account for ordinary comprehension of continuous discourse will have to incorporate heuristic strategies, and to assume that higher-level processes (syntax, semantics, context) are involved from the very beginning. One such strategy, based on this idea of simultaneous and parallel processing, is redundancy utilisation. It may be defined as the listener's capacity to make systematic hypotheses about the content and structure of the message. This capacity has its basis in the linguistic and pragmatic competence of the listener, and it enables him to construct projective hypotheses from earlier information and to evaluate these hypotheses utilising subsequent data. The listener's knowledge of communication situations tells him what is likely to be said in the given situation. Further, his knowledge of syntax tells him what to expect if, for example, he hears a question word at the beginning of a sentence. Similarly, he knows the phonological rules concerning permissible sound sequences, and these rules facilitate word recognition by introducing redundancy on the phonological level, and so on. The important thing is that the utilisation of redundancies at different levels of language is a simultaneous and parallel process, i.e., that comprehension performance is integrative.

The constructivist view has several implications. First of all, it means that processing is facilitated by the linguistic and pragmatic organisation of the message, together with its presentation

in context. The more explicit and clear the structure, the greater the amount of redundancy in the message, and the easier the comprehension task. Similarly, if the utterance is tied to a sufficient context, it is likely that comprehension will be enhanced through expectations and a psychological set that effectively guides the listener's attention[1]. If organisation of the material and its presentation in context are important aids to language processing by natives, they are all the more important for the foreign language learner, whose mastery of the new linguistic system may still be relatively incomplete. Native speakers talking to learners are intuitively aware of this, and tend to avoid speaking in ways that would make the organisation of their message less evident. Pragmatic organisation of the material is particularly important for the processing that takes place in a language learning situation, since it increases the learner's opportunities for utilising redundancies that are not language-specific.

Secondly, the constructivist view implies that memory, especially short-term storage, is an essential part of comprehension. Short-term "working" memory is known to have a limited capacity, but since it operates on the principle of hierarchical recoding ("chunking"), its capacity can be extended. The listener's generative competence, which allows redundancy utilisation, also accounts for his ability to recode verbal material for retention in the working memory. The retention of a certain amount of previous context is necessary for the generation of anticipatory hypotheses. Similarly, the "right-to-left" clarification of unclear or ambiguous points in the message also requires the storage of fairly extensive sequences. As the native listener masters the categories of his language, he is able to "chunk" the message according to these categories. As a result, he is able to hold large amounts of context in his working memory, and to utilise this context while anticipating or clarifying parts of the message. The L2 learner, however, has an imperfectly developed mastery of the categories of the language, which

---

[1] Psycholinguistic research on the effects of organisation and context on comprehension, memory, and learning has been reviewed by Oller (1972) and Kohonen (1975).

presumably reduces his ability to handle data in short-term storage. Some experiments that suggest a connection between L2 skill and short-term memory are described by Lado (1965; 1970) and Harris (1970). It has been shown, for example, that short-term memory capacity is greater for L1 than for L2 material; that L2 sentence memory span is affected by certain grammatical variables to a larger extent than the L1 span; and that L2 short-term memory tests correlate reasonably well with typical multiple-choice tests of listening comprehension and grammar. The likely explanation for these L2 results is that the learner lacks the native speaker's full capacity for effective recoding. Therefore, his capacity to handle L2 data in the working memory is limited, and this is reflected in both comprehension and production.

Looking at the active processes of comprehension and redundancy utilisation from another angle, a third implication of the constructivist view is that the native listener is only partly bound by the properties of the signal he receives. In ordinary communication, the listener must actively construct the message for himself. One reason for this is that spoken discourse is often highly fragmentary and elliptical. The clarity of the message may also be reduced because of psychological or even physiological performance limitations either in the speaker or in the listener. The channel of communication may carry a considerable amount of noise, competing messages, and so on. The presence of some kind of disturbance is the rule rather than the exception in normal communication. Under such circumstances, a purely "passive" comprehension system, entirely dependent on the stimulus, would in fact be an impossibility.

The ability of the competent native listener to process non-ideal messages was efficiently demonstrated by the early experiments with acoustic speech distortion. The experimenters used filtering, amplitude selection, time compression, noise addition, and a variety of other techniques. The general finding was that continuous speech is highly resistant to distortion (see Miller 1951). The explanation for this was found in the high redundancy of language, i.e. the multiplicity of cues present in the message. According to the constructivist view, another equally crucial factor is the reconstructive power of the native listener's

competence. In the case of the L2 learner, this creative redundancy utilisation capacity is normally more restricted.

3. Two reduced redundancy tests of listening comprehension

3.1. General remarks

The central point in the constructivist view of comprehension is thus the active contribution of the listener. He puts the message together for himself, working simultaneously at all levels of language structure, and utilising whatever linguistic or extra-linguistic cues there may be in the message, in order to arrive at the meaning as quickly as possible. Such an integrative capacity for redundancy utilisation can even be thought of as a central factor underlying much of language use (cf. Holzman 1967; Oller 1976). According to this view, a convenient way to get an overall estimate of the L2 learner's skills is to present him with messages that do not contain all the information they normally carry. The result - a reduced redundancy test - is integrative in much the same way as language use itself is integrative. In such tests, the learner needs to mobilise his total awareness of the L2 linguistic and pragmatic structure. He uses that structure in making hypotheses about the content of the message. His memory capacity is also involved, since memory is necessary for both the making and verification of these hypotheses.

3.2. The "Noise Test"

In a language testing context, the first systematic experiments with noise addition were carried out at Indiana University by Spolsky et al. (1968; see also Spolsky 1971; 1975). At Lund University, the technique has been used by Johansson (1972; 1973). Further investigations of its characteristics as a FL test include Gradman (1973) and Gradman & Spolsky (1975). Some new development in the contextualisation of these tests have been reported by Seliger & Whiteson (1975), Whiteson (1975), and Gaies et al. (1977).

The term "noise test" has come to refer to the reduced redundancy listening tests employed by these testers. In this type of test, the subject hears sentences mixed with white noise (a hissing sound in which the frequencies of a wide frequency

range are equally represented). The sentences are usually unconnected. The task of the testee is to reproduce the sentences in writing as accurately as possible. Thus, the noise test is essentially a dictation test where the entire message must be processed for both comprehension and verbatim reproduction, and where the main emphasis is on intensive listening. In its use of external distortion, it is related to those tests of (usually extensive) comprehension where authentic materials (street interviews against traffic noise, for example) are used, but the noise test represents an attempt to control the level of redundancy reduction. The testee's proficiency can be stated either in terms of the signal-to-noise ratio with which he achieves a given level of accuracy, or in terms of his comprehension score at a fixed level of distortion.

The preliminary experiments by Spolsky et al. were concerned mainly with the practical arrangements required by the technique. It was found that an adequate scoring system for the dictations can be developed, and that the noise can be added without too much difficulty, although the use of various signal-to-noise ratios demands electronic compressing of the signal. In addition to the noise-plus-dictation technique, a multiple-choice variant has been used (Gradman 1973), in which the testee selects his responses from 5 alternative sentences. In general, the noise test appeared reasonably viable. However, the technical aspects of its construction can be complicated, and the continuous presence of disturbing hiss may be found objectionable by some testees.

Spolsky (1971) and Johansson (1972) agree that the test appears reliable even in its dictation form. The multiple-choice version used by Gradman had a Kuder-Richardson reliability index of .68 (N=25). When the test was given to over six hundred non-native speakers of English, its reliability reached .92. The multiple-choice version was found to correlate at over .80 with the dictation version, indicating that there would be some grounds for using the administratively more efficient and objective variant instead of the subjectively scored dictation (Gradman & Spolsky 1975).

Most of the discussion around noise tests of the above type

has centred on the validity of the technique. Its construct validity as a reduced redundancy test is fairly generally accepted. Some differences of opinion seem to exist about empirical validity, estimated through correlations between noise tests and other measures of L2 proficiency. The final form of Spolsky's original (1968) test correlated at .66 (Spearman's rho) with a diagnostic test of listening comprehension, at .66 with an objective discrete-point grammar test, and at .51 with an essay (N = 61). The continuation of these experiments reported by Gradman (1973) and Gradman & Spolsky (1975) produced reasonably high coefficients both with oral interviews (.79 and .69; N = 25,26) and with the totals of the TOEFL battery (.75 and .66, respectively). With a test modelled on Spolsky's original, Whiteson (1972) obtained a .54 correlation between noise test results and a proficiency test developed at Cambridge University (N = 12). The contextualised type of noise test (Seliger and Whiteson 1975) correlated at .69 (N = 65) with a test based on the Cambridge Proficiency Exam. Johansson (1972; 1973) reports a .52 correlation with what is termed a "test of general English proficiency" (1972:399) - a written text with a number of points where the testee chooses the syntactically right one from 2-3 alternative formulations.

The interpretations of these figures differ considerably. Spolsky and Gradman think that the noise test is a promising measure of global L2 proficiency. Gradman (1973) hopes that it may be able to replace the functionally oriented but subjective interview. Johansson, however, comes to the conclusion that the noise test is not a valid measure of L2 overall skill.

Although Johansson's test of general English proficiency may not be entirely adequate as a criterion, the correlations do not on the whole seem to warrant the conclusion that a valid estimate of L2 skill could be based on noise tests alone. At best, the noise test seems to account for some 50 per cent of the variance in a traditional proficiency battery. It is also fairly easy to agree with Johansson that the noise test - as rather a special kind of listening test - is likely to emphasise psychological and even physiological factors such as hearing ability and ability to disregard external disturbance when concentrating on a task.

A second point of interest related to validity is the relationship between the noise test and dictation as integrative listening tests. Although Johansson's low correlation between the

noise test and a test of listening comprehension (.20) in part results from the low discrimination power of the former, the fact remains that dictation without noise correlated considerably better with listening comprehension (.83; N = 26). The correlation of .93 (N = 71) obtained by Gradman and Spolsky (1975) between dictation with heavy distortion and dictation with minimal distortion (S/N = 50 dB) also indicates that dictation could be used alone, much as Johansson suggests. Gradman and Spolsky defend the use of noise by saying that mistakes at different levels of distortion may be different, and that further research into the use of noise could give useful information on how linguistic units relate to levels of redundancy reduction. These questions are certainly interesting for interlanguage analysis. If one wants to use dictation methods in testing L2 skills, the traditional undistorted exercise would, however, seem simpler and more straightforward, especially as it also contains a strong redundancy utilisation element (cf. Oller and Streif 1975). Recently Gaies, Gradman, and Spolsky (1977:56) have in fact come round to Johansson's (1972) view, and state that "the use of background noise seems to have had little effect on the measurement of overall English proficiency" for their students.

The fact that isolated sentences have normally been used in noise tests also diminishes their validity as integrative measures of listening proficiency. Contextualisation of these tests (the use of a continuous text in a lifelike situation, e.g. a conversation between two airline passengers with jet whine in the background) is certainly a desirable improvement (Seliger and Whiteson 1975; cf. also Gaies et al. 1977). However, it seems a little dubious whether even this increased "realism" can avert hostile attitudes in students taking such tests on repeated occasions. And since the addition of noise seems to make relatively little difference in the results, its value as a systematic redundancy reduction tool in intensive listening tests remains unproven.

3.2. Aural cloze

The cloze procedure, originated by Taylor (1953) as a readability measure and subsequently used by native and foreign language testers as a reading comprehension test, also lends itself fairly

easily to reduced redundancy testing of listening skills. One possible method is to use an unmutilated listening passage together with a cloze-mutilated transcript on which the testees mark their responses after listening (Oakeshott-Taylor 1976; cf. also Nygård & Nordström 1976). The filling of the gaps may also be done as a speeded test to simulate the real time limits of spoken communication. Another variant of this technique is Johansson's "partial dictation" (1973), in which pauses are inserted at suitable places after each gap for the student to fill in the missing section of the aural and written message. These procedures are essentially written cloze made easier by the preceding or partly simultaneous presence of the original information in aural form.

A somewhat more aural technique is the presentation of a listening passage or dialogue that is interrupted at certain times, when the testee is asked to choose the continuation from a set of written alternatives. This produces a test of the ability to utilise previous context in "predictive listening" (Hughes 1974). It may be a measure of both extensive listening, including inference, and more detailed points of language structure and usage. The problems inherent in item construction are perhaps compensated for by quick and reliable scoring.

The original idea of the cloze procedure as a systematic reduction of redundancy through statistical deletion can also be applied to taped aural messages. A given stretch (e.g. one second) of acoustic information may be excised at regular intervals and replaced by silence. Another possibility is to base the deletion on linguistic units, words or longer segments. Although the latter procedure may be criticised on phonetic grounds, the segmenting can, however, be done relatively easily and with an accuracy sufficient for measuring the testees' expectancies of the linguistic features involved.

Both the noise test and aural cloze of the word-deletion type can probably be characterised as intensive listening tests. They are integrative in the sense that both continuous acoustic distortion, and the random deletion of words, result in an even spread of redundancy reduction over different linguistic features. Short-term memory plays a similar role in both types of test, and both of them require fairly quick reaction in the use of language, as is appropriate in a test connected with oral communication. Verbatim

reproduction, however, is more important in the dictation-type noise test: if the requirement of verbatim reproduction is dropped, the scoring of semantically acceptable alternatives soon becomes problematic. Aural cloze is more easily scored according to contextual criteria, and may thus give more emphasis to adequate comprehension. Although aural cloze is not very far removed from dictation in the sense that it requires the student to take down parts of the message, it may demand a more active and creative utilisation of redundancies than the noise test does.

Written cloze has been subjected to extensive research, but relatively little can be found about aural cloze as a testing technique, although its use was suggested by Taylor as early as the fifties (Taylor 1956).

Dickens and Williams (1964) describe an experiment in which two aural cloze tests were given to native speakers of English (N = 127, 126). The tests, in which the reduction of redundancy consisted in the deletion of one word every 5 seconds, had split-half reliabilities of .80 and .70, and correlated with each other at .73. The scores were found to correlate at .52 and .49 with a general test of English language ability. These correlations were almost identical with those obtained between the language ability test and two traditional tests of listening comprehension. However, according to the authors, aural cloze appeared more reliable. The reliability and validity of aural cloze as a foreign-language test were further investigated by Gregory-Panopoulos (1966), whose test contained 126 items (deletions of every 5th word from parts of a lecture), with the maximum of 4 deletions per sentence. Verbatim scoring was used with some exceptions. Test-retest reliability was slightly over .90 in the two adult subject groups (N = 43, 47), and the test was found to correlate at .66 with a standardised listening comprehension test, and at .69 with the California Reading Test. The author considers cloze to be a more direct and reliable measure of foreign-language listening comprehension than the traditional multiple-choice tests. Nutter's (1974) study of the effects of different deletion patterns and other technical variations in the procedure indicated, not surprisingly, that reliabilities and difficulty levels are affected by such factors as passage type, different methods of presentation, and deletion patterns. Templeton's (1977) test, given to a group of 39 adult

foreign students of English, had a reliability of over .90. It correlated best with the aural section of a proficiency battery (.80), while the lowest correlation was obtained with the vocabulary section (.54).

The aural cloze tests mentioned above have displayed quite a remarkable reliability. This may in part be connected with the fact that it is fairly easy to make a cloze test that contains a large number of items. Their construct validity is again based on the ideas of language redundancy and constructive comprehension. As far as empirical validity is concerned, Templeton's results, for example, seem promising. More work should, however, be carried out on the question of what exactly is involved in a test of this type, i.e. whether it can be called a test of general L2 proficiency, a listening comprehension test, a vocabulary quiz, or, in Templeton's phrase (1977:298), a test of "spotting the bleep". The next section of this paper deals with some of these questions in a very preliminary way.

4. Aural cloze with Finnish learners: some tentative results

Some preliminary experiments with aural cloze were carried out in 1976-7 at the Department of English, University of Turku, in order to find out whether an intensive listening test could be developed that would have a stronger theoretical basis than the completion tasks traditionally used to measure listening for details. It was felt that in many cases the completion items in existing listening comprehension exercises and tests were written because it had been difficult to construct enough sensible multiple-choice questions for the exercise to be usable or the test to be reliable. The idea of redundancy utilisation seemed to offer a suitable background for the development of such intensive listening material. The skill of intensive listening was loosely defined as the ability to concentrate on the language used to express the content of the passage. The task of the student in the test can be said to include most of Hughes' (1974) categories of listening skills (at least predictive, retrospective, constructive, and even inferential listening), but the emphasis here was on language structure. Thus the aim of the test was not

to measure how well the student could follow the general lines of argument in an aural message, or how well he could extract information from the passage and evaluate this information, although this wider type of comprehension is not doubt involved in intensive listening as well. An integrative type of language proficiency test was aimed at, but it was assumed that results would be affected by the fact that the test was specifically a listening test.

Accordingly, a number of aural cloze tests based on the deletion of words from taped monologues (originally either scripted radio talks or written passages modified for oral presentation) were administered to first-year students of English, and a group of native British secondary school students. So far, detailed results are available for the first of the experimental EFL tests. In the following some general observations about this test are given.

The test consisted of 72 items, with the omission of every 10th word as the deletion principle. The deleted words were replaced by splicing in a quiet electronic signal of constant length (1 sec.). The following extract illustrates the type of text used[1]:

> ... Well, I'm a professional writer, and _____ I was younger I thought a typewriter would be _____. I even thought it was necessary, and that editors and _____ would expect anything sent to them to be typewritten. _____ I bought myself a typewriter and taught myself to _____. And for some years I typed away busily. But _____ didn't enjoy typing. I happen to enjoy the act of _____ ...
>
> (adapted from Spencer, D.H., English for Proficiency, OUP 1963.)

In this experiment, unlike the subsequent ones, the subjects were not allowed to listen to the mutilated passage beforehand. Their task was to reconstruct the passage on the basis of only one hearing, and to write down their completions on an answer sheet either during listening or during pauses that were inserted at

---

[1] The texts selected contained a considerable amount of redundancy in the form of repetition, as can be seen from the extract. In some texts, the amount of redundancy was increased by shortening the sentences (converting embeddings into short main clauses, for example), or by inserting reformulations of words, etc. - Occasionally the context between two gaps was shortened or lengthened by one word owing to such phonetic reasons as reductions in weak forms.

suitable syntactic boundaries, mostly between sentences.

The response data were processed using the CLOZE computer program (see Kohonen & Salmela 1977), and the subjects' completions were evaluated according to both verbatim and contextual criteria. In the latter scoring, which was done by a native English speaker, a completion was counted as correct if it fitted the context up to the end of the last sentence heard. An item analysis was carried out using the OPSAM program (Mikkonen & Mikkonen 1971). Table 1 gives some of the item analysis results.

TABLE 1. ITEM ANALYSIS OF AN EXPERIMENTAL AURAL CLOZE TEST

No of items = 72
N = 55

	1	2	3	4
verbatim scoring	40 %	.80	56 (78 %)	.82
contextual scoring	54 %	.84	65 (90 %)	.85

1 = average solution percentage; 2 = reliability (internal consistency) of the entire 72-item test; 3 = number and (percentage) of items with an item correlation $\geq .00$; 4 = reliability of test with items mentioned in 3.

As was expected, the test in this form (only one hearing) proved rather difficult. An interesting result was the relatively small difference between the average solution percentages on the two scorings. This was due to the scoring criteria used in 1976, since paraphrases of several words were not accepted in this contextual scoring. The criteria have later been modified, and this will probably increase the difference between verbatim and contextual percentages. A second listening, employed in later testings, also makes the task of the testee somewhat easier. It may be seen that the reliability of the test was relatively high even before any items had been removed on the basis of the

item correlations[1]. Contextual scoring is seen to be statistically more effective, besides being intuitively a more proper way to assess foreign language cloze responses (cf. Kohonen 1976, whose findings in an EFL written cloze test were similar). The general observation from the item analysis is that the experimental aural cloze test proved to be a fairly reliable instrument, and it is probable that the changes introduced in the scoring criteria will further diminish the number of discarded items.

The previous research reported above seems to indicate that reduced redundancy tests of this type have reasonable empirical validity. In order to see how the test performed in that respect, the scores were correlated with the subjects' scores in the 1976 Joint Entrance Exam (arranged by Turku together with three other university Departments of English).

TABLE 2. CORRELATIONS BETWEEN AURAL CLOZE AND THE 1976 JOINT ENTRANCE EXAM

N = 48

	1	2	3	4	5	6	7
verbatim	.15	.49+	.39+	.17	.28	.11	.49+
contextual	.10	.48+	.40+	.24	.34	.22	.52+

1 = lecture comprehension; 2 = listening comprehension; 3 = reading comprehension; 4 = vocabulary; 5 = grammar; 6 = "verbal reasoning" (linguistic inferences from artificial language data); 7 = total of sections 1-5. Asterisk = significant beyond the 99 per cent level of confidence.

[1] In the analysis, items can be discarded according to a pre-determined level of item correlation. Normally, zero correlation is a suitable cut-off point for small populations such as the present group. - If one employs the more stringent criterion that a "good" item must correlate with the total score at the 95 per cent level of confidence, about half of the items (31,ie. 51 %) are functioning well with contextual scoring. The corresponding figure for verbatim scoring is 26 (= 36 %). This reflects the typical characteristic of cloze tests that they contain relatively many items that are either too easy or too difficult, and the fact that contextual scoring functions better than verbatim scoring.

As may be seen from the table, the product-moment correlations are not quite as high as in some of the experiments reported earlier; however, some of them are statistically significant beyond the 1 per cent risk level, and the correlation between aural cloze and the English skills total (sections 1-5 of the Entrance Exam) seems high enough to warrant further use of the test. Since the test population represented the top candidates in the entrance test, the variance in their entrance test results is relatively small. This is thus one factor that in fact prevents the correlation from attaining a particularly high value.

The correlations in Table 2 must be treated with the usual caution, but perhaps they permit some speculations about what kinds of skill are involved in aural cloze. Contextually scored aural cloze correlates best with aural comprehension. This section contained several items testing "predictive listening" - multiple-choice completions of dialogues, testing the utilisation of preceding context both extensively and from the point of view of usage and idioms. The correlation between contextual aural cloze and lecture comprehension, on the other hand, is the lowest of the set. The lecture questions concentrated on measuring the acquisition and even application of information contained in a taped lecture. These correlations may perhaps be taken as an indication that aural cloze is more concerned with intensive than extensive listening, and that it is more a test of language structure in a wide sense than a test of following the argumentation in an aural message. The differences between the rest of the correlations are too small to justify proper comments. However, it may be a sign of the integrative nature of aural cloze that its correlations with reading and listening comprehension - both of them integrative tests - are somewhat higher than its correlations with the discrete-point grammar and vocabulary sections. Possibly one could also speculate that the mastery of grammar, and of language structure in general, is at least as important as the mere knowledge of vocabulary in tests of this type.

No doubt the reduced redundancy tests described above - the noise test and aural cloze - involve a considerable element specific to listening. It is thus difficult to say whether they would suffice alone as measures of L2 proficiency, even if a suitable combination of statistical factors were to produce high coefficients

of validity. As subtests in proficiency batteries they seem to be useful. Gradman and Spolsky (1975), for example, think that the noise test can cover an unattended area between functionally oriented interviews and discrete-point measures of proficiency. As far as can be told from the present preliminary results, aural cloze may be even more interesting as an integrative but structured test, since it contains a considerable element of creativity, and may thus serve as a more effective indication of the learner's fluency than the noise test.

The role of "local redundancy" (Carroll 1972) is obviously important in aural (and written) cloze: most of the items tend to measure the utilisation of short-range constraints. However, this is by no means an unimportant aim of language learning (cf. Enkvist and Kohonen 1977), and exercises or tests of this intensive ability therefore seem justified. But the fact that local redundancy is central in many cloze tests means that they can hardly be employed as the test of L2 comprehension. The more extensive aspects of listening, such as separating the essential information from unimportant detail (Hughes' "redundancy listening") are probably best measured by multiple-choice and related techniques.

One disadvantage of aural cloze of the above type is that only restricted kinds of material can be used. Another technical difficulty is that the mutilation of an aural message introduces the somewhat irrelevant task of "spotting the bleeps", and keeping the exact locations of the omissions in mind. This can be remedied in part by telling the subjects that they can write down not only the missing item but also some of its immediate context.

On the whole, the completions seem to produce fairly interesting insights into how learners process spoken text. The fact that it is the weaker students who tend to get mixed up is explained by their imperfect redundancy utilisation capacity: making hypotheses about the content of the gaps takes so long that inadequate time is left for rehearsal. As a result, parts of the information vanish from the working memory, and the content and form of the message are inadequately processed. The more fluent learner, on the other hand, is capable of quick recognition of the linguistic features in the message; his fluency also makes for efficient hypothesis-generation and quick recoding. This leaves more time for rehearsal, and the reproduction of the content as well as the form of the message is easier.

The face validity of the reduced redundancy tests mentioned above should also be briefly commented on, as it may be the most important thing to the "layman" taking the test. The intentional presence of distorting noise is hardly likely to make the test "seem right" to the testees. As far as aural cloze is concerned, similar doubts are expressed by some testees about the relevance of gap-filling – an indication that the connection between redundancy utilisation, as measured by cloze, and communication, may not be immediately recognisable. On the other hand, the majority of students in the experiment described above, and in the subsequent testings, considered the test interesting and valid in the sense that it requires imaginativeness and a fluent general mastery of the language.

It is of course desirable that the tests given to learners should also have some instructional value. The idea of redundancy utilisation is clearly not confined to tests only. It is inherent, for example, in the use of authentic materials for extensive listening practice. A redundancy reduction element is contained in all exercises where normal rates of delivery are used, where there are several speakers engaged in spontaneous conversation with normal amounts of background noise, etc. Anticipatory listening, a central feature in reduced redundancy tests, can be practised in class by interrupting the presentation of the exercise tape at certain places and asking students to suggest and evaluate possible continuations. The kind of aural cloze that was described above can probably be used at advanced levels as an exercise for increasing the learners' L2 processing capacity. According to Godfrey (1977:118), such an increase would follow from "pushing the learner to do more with the utterance", i.e. involving him in structured active deep-level processing. Given the time limits inherent in spoken communication, this would force the learner to make full use of his chunking and working memory capacities. He would thus in fact be creating "extra" processing time, which could be channeled into rehearsal (for better retention and recall) and further recoding (e.g. relating discourse segments to each other). In aural cloze the listener is actively engaged in construction that is typical of real comprehension, and the generation of hypotheses about missing items involves deep-level processing of the message. The fact that the elusive spoken message must be attended to in depth may be one reason why students normally find aural cloze exercises fairly interesting and instructive[1].

---

[1]I am grateful to Viljo Kohonen and Keith Battarbee for comments.

REFERENCES:

Bormuth, J. 1962. Cloze Tests as Measures of Readability and Comprehension Ability. Diss., Indiana University.

Carroll, J.B. 1972. Defining language comprehension: some speculations. In Freedle, R.O. & Carroll, J.B. (Eds.), 1-29.

Dickens, M. & Williams, F. 1964. An experimental application of "cloze" procedure and attitude measures to listening comprehension. Speech Monographs 31, 103-8.

Enkvist, N.E. & Kohonen, V. 1977. Cloze testing, some theoretical and practical aspects. In Zettersten, A. (Ed.), 11-42.

Enkvist, N.E. & Kohonen, V. (Eds.). To appear 1978. Text Linguistics, Cognitive Theory and Language Teaching. Turku: AFinLA.

Freedle, R.O. & Carroll, J.B. (Eds.) 1972. Language Comprehension and the Acquisition of Knowledege. Washington, D.C.: V.H. Winston & Sons.

Gaies, S., Gradman, H. & Spolsky, B. 1977. Toward the measurement of functional proficiency: contextualization of the noise test. TESOL Quarterly 11:1, 51-7.

Godfrey, D. 1977. Listening instruction and practice for advanced second language students. Language Learning 27:1, 109-22.

Gradman, H.L. 1973. Reduced redundancy testing: a reconsideration. In O'Brien, M. (Ed.) 41-8.

Gradman, H.L. & Spolsky, B. 1975. Reduced redundancy testing: a progress report. In Jones & Spolsky (Eds.), 59-70.

Gregory-Panopoulos, J.F. 1966. An Experimental Application of "Cloze" Procedure as a Diagnostic Test of Listening Comprehension among Foreign Students. Diss. University of Southern California.

Harris, D.P. 1970. Report on an experimental group-administered memory span test. TESOL Quarterly 4:3, 203-13.

Holzman,P. 1967. English language proficiency testing and the individual. Selected Conference Papers of the Association of Teachers of English as a Second Language 76-84. Los Altos, California: Language Research Associates Press.

Hughes, G. 1974. Aspects of listening comprehension. Audio-Visual Language Journal 12:2, 75-9.

Johansson, S. 1972. Controlled distortion as a language testing tool. In Qvistgaard et al. (Eds.), 397-411.

Johansson, S. 1973. An evaluation of the noise test. IRAL 11, 107-34.

Jones, R.L. & Spolsky, B. (Eds.) 1975. Testing Language Proficiency. Center for Applied Linguistics.

Kohonen, V. 1976a. Kontekstin osuudesta vieraankielen oppimisessa ja kielitaidon testaamisessa. In Nummenmaa (Ed.), 23-43.

Kohonen, V. 1976b. Cloze-testien ongelmia testianalyysin valossa. In Kohonen, V. (Ed.), 31-54.

Kohonen, V. (Ed.) 1976c. Integratiivisen kielitaidon mittaamisesta cloze-testien avulla: teoriaa ja sovelluksia. Turku: AFinLA.

Lado, R. 1965. Memory span as a factor in second language learning. IRAL, Vol. III, 123-9.

Lado, R. 1970. Language, thought and memory in language teaching: a thought view. The Modern Language Journal LIV: 8, 580-5.

Marslen-Wilson,W. 1976. Linguistic descriptions and psychological assumptions in the study of sentence perception. In Wales & Walker (Eds.), 203-29.

Massaro, D.W. 1975. Language and information processing. In Massaro (Ed.) 1975, 3-28.

Massaro, D.W. (Ed.) 1975. Understanding Language. New York: Academic Press.

Miller, G.A. 1951. Language and Communication. New York: McGraw-Hill.

Miller, G.A. 1956. The magical number seven plus or minus two. Psychological Review 63, 81-97.

Mikkonen, V. & Mikkonen, J. 1971. OPSAM: opintosaavutusten mittaus. Helsinki: Tammi.

Neisser, U. 1966. Cognitive Psychology. New York: Appleton-Century-Crofts.

Nummenmaa, L. (Ed.) 1976. Kieli, konteksti ja tilanne. AFinLAn syyssymposiumi 1975. Helsinki:AFinLA.

Nutter, B. 1974. Presentation Methods, Deletion Patterns, and Passage Types for Use with Aural Cloze. Diss. University of Arizona. Dissertation Abstracts 1974, 35, 945-A.

Nygård, A. & Nordström, H. 1976. Två cloze test i svenskt inträdesprov till Åbo Akademi. In Kohonen, V. (Ed.), 73-87.

Oakeshott-Taylor, J. 1976. Cloze procedure and foreign language listening skills. L.A.U.T. Series

O'Brien, M. (Ed.) 1973. Testing in Second Language Teaching: New Dimensions. Dublin: ATESOL.

Oller, J.W. 1972. Contrastive analysis, difficulty, and predictability. Foreign Language Annals 6, 95-106.

Oller, J.W. 1976. Evidence for a general language proficiency factor: an expectancy grammar. Die Neueren Sprachen 2/76, 165-74.

Oller, J.W. & Richards, J.C. (Eds.) 1975. Focus on the Learner: Pragmatic Perspectives for the Language Teacher. Newbury House.

Oller, J.W. & Streif, V. 1975. Dictation: a test of grammar-based expectancies. English Language Teaching Journal XXX:1, 25-36.

Perren, G.E. & Trim, J.L.M. (Eds.) 1971. Applications of Linguistics. Cambridge: Cambridge University Press.

Pimsleur, P. & Quinn, T. (Eds.) 1975. The Psychology of Second Language Learning. Cambridge: Cambridge University Press.

Qvistgaard, J. et al. (Eds.) 1972. AILA Proceedings 1972, Vol. III. Heidelberg.

Rivers, W.M. 1975. Linguistic and psychological factors in speech perception and their implications for teaching materials. In Pimsleur & Quinn (Eds.), 123-34.

Seliger, H.W. & Whiteson, V. 1975. Contextualizing laboratory administered comprehension tests. System, 3:1, 10-15.

Spolsky, B. et al. 1968. Preliminary studies in the development of techniques for testing overall second language proficiency. Language Learning Special Issue No 3, 79-101.

Spolsky, B. 1971. Reduced redundancy as a language testing tool. In Perren & Trim (Eds.), 381-90.

Spolsky, B. 1975. What does it mean to know a language; or how do you get someone to perform his competence? In Oller & Richards (Eds.), 164-76.

Taylor, W. 1953. 'Cloze procedure': a new tool for measuring Readability. Journalism Quarterly, Fall 1953, 415-33.

Taylor, W. 1957. Cloze readability scores as indices of individual differences in comprehension and aptitude. Journal of Applied Psychology XL 19-26.

Templeton, H. 1977. A new technique for measuring listening comprehension. English Language Teaching Journal XXXI:4, 292-9.

Wales, R.J. & Walker, E. 1976. New Approaches to Language Mechanisms. Amsterdam etc.: North Holland Publishing Co.

Wanner, E. 1974. On Remembering, Forgetting, and Understanding Sentences. The Hague: Mouton.

Whiteson, V. 1972. The correlation of auditory comprehension with general language proficiency. Audio-Visual Language Journal 10:2, 89-91.

Whiteson, V. 1975. An integrative approach to the 'noise test'. Audio-Visual Language Journal 13:1, 17-8.

Wold, A.H.1976. Decoding Strategies, Word Openness and the Dimension of Time in Oral Language. Oslo: University of Oslo.

Zettersten, A. (Ed.) 1977. Papers on English Language Testing in Scandinavia. Anglica et Americana 1, Dept. of English, University of Copenhagen.